New Data Sources – Accessibility and Use

Frauke Kreuter

JPSM – Uni Mannheim – IAB

Paris 6/25/2019

AAPOR

AMERICAN ASSOCIATION FOR PUBLIC OPINION RESEARCH

AAPOR Report on Big Data

AAPOR Big Data Task Force

February 12, 2015

Prepared for AAPOR Council by the Task Force, with Task Force members including:

Lilli Japec, Co-Chair, Statistics Sweden
Frauke Kreuter, Co-Chair, JPSM at the U. of Maryland, U. of Mannheim & IAB
Marcus Berg, Stockholm University
Paul Biemer, RTI International
Paul Decker, Mathematica Policy Research
Cliff Lampe, School of Information at the University of Michigan
Julia Lane, American Institutes for Research
Cathy O'Neil, Johnson Research Labs
Abe Usher, HumanGeo Group

Acknowledgement: We are grateful for comments, feedback and editorial help from Eran Ben-Porath, Jason McMillan, and the AAPOR council members.

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

REPORT

INNOVATIONS IN FEDERAL STATISTICS

Combining Data Sources While Protecting Privacy

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

FEDERAL STATISTICS, MULTIPLE DATA SOURCES, AND PRIVACY PROTECTION

Next Steps

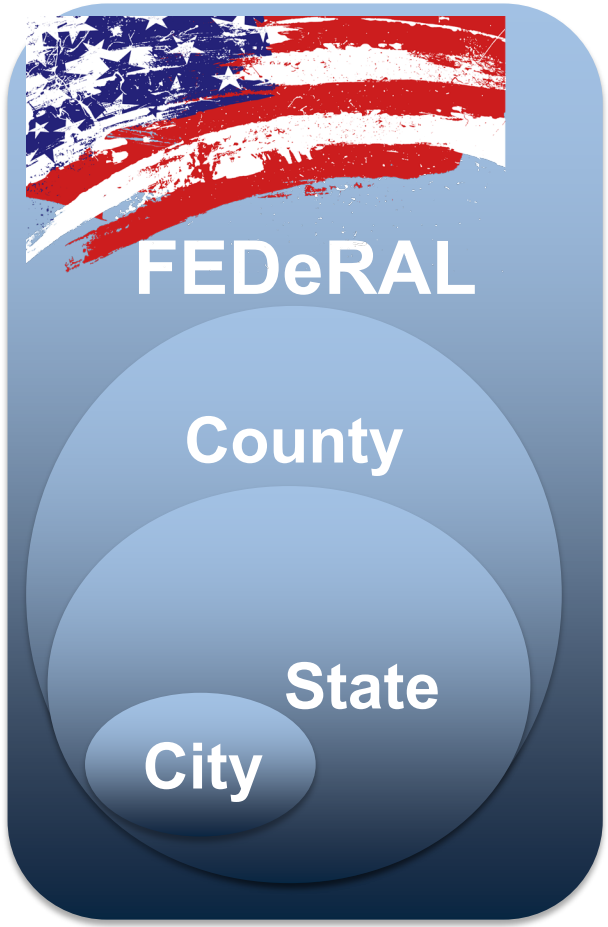
SPONSORED BY THE

Federal Ministry of Education and Research

AUFSTIEG DURCH BILDUNG
OFFENE HOCHSCHULEN

The project on which this report is based was funded by the Federal Ministry of Education and Research under the number [16OH22064]. Responsibility for the contents of this publication lies with the author.

~40 agencies from city, state, county and federal agencies



HEALTHY CHICAGO
CHICAGO DEPARTMENT OF PUBLIC HEALTH

MECKLENBURG COUNTY NORTH CAROLINA

NEW YORK CITY HOUSING AUTHORITY

CHICAGO POLICE

NEW YORK STATE OF OPPORTUNITY.
Office of Information Technology Services

Virginia Department of Behavioral Health & Developmental Services

USDA

City of Milwaukee

Illinois Department of Corrections

THE CITY OF SOUTH BEND, INDIANA
PEACE
1865

IDES
ILLINOIS DEPARTMENT OF EMPLOYMENT SECURITY

Allegheny County Department of Human Services

UNITED STATES GOVERNMENT ACCOUNTABILITY OFFICE

U.S. DEPARTMENT OF HOUSING AND URBAN DEVELOPMENT

COUNTY of DANE WISCONSIN 1839

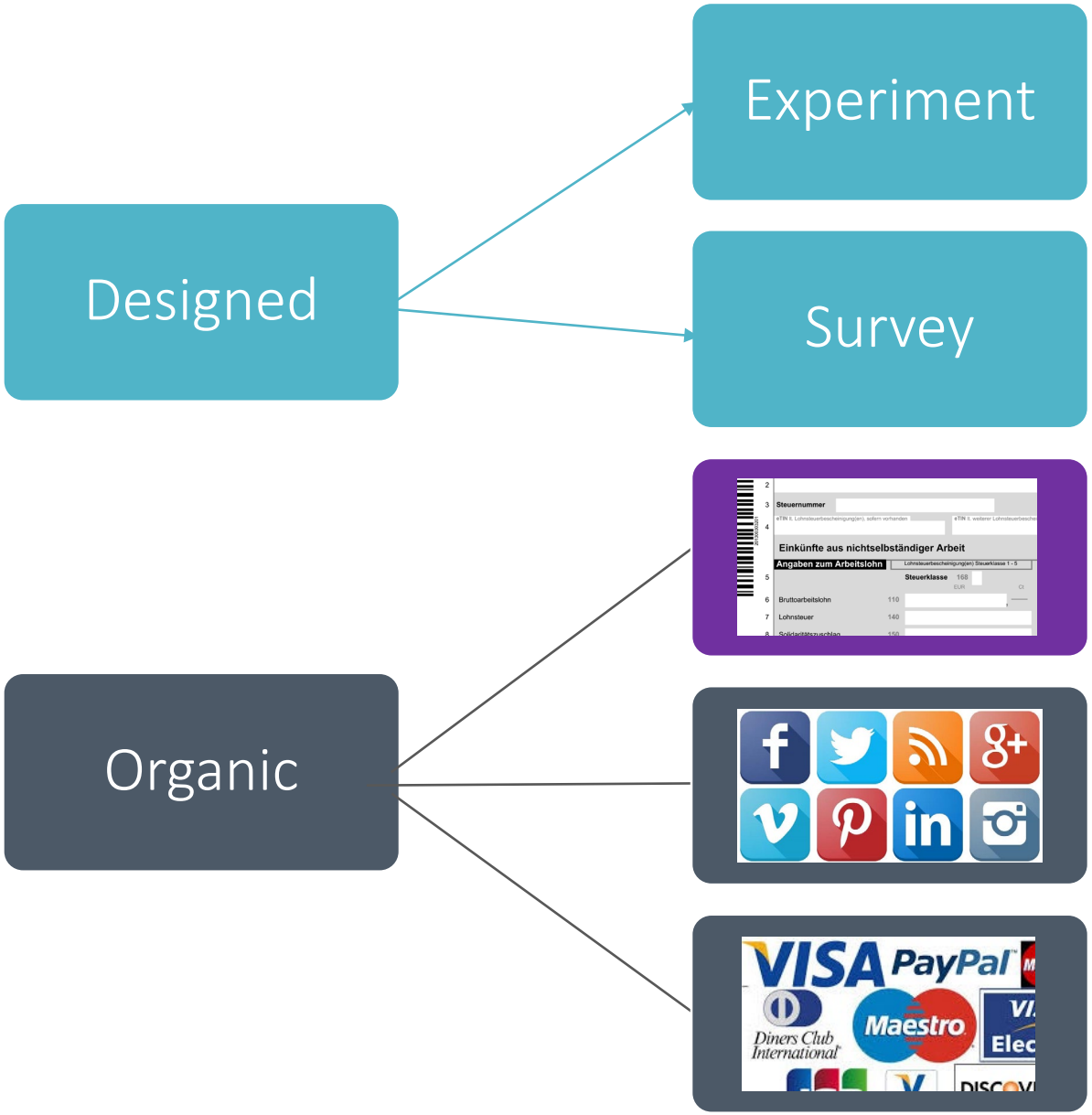
NYC .gov always open

NEW YORK STATE Office of Alcoholism and Substance Abuse Services

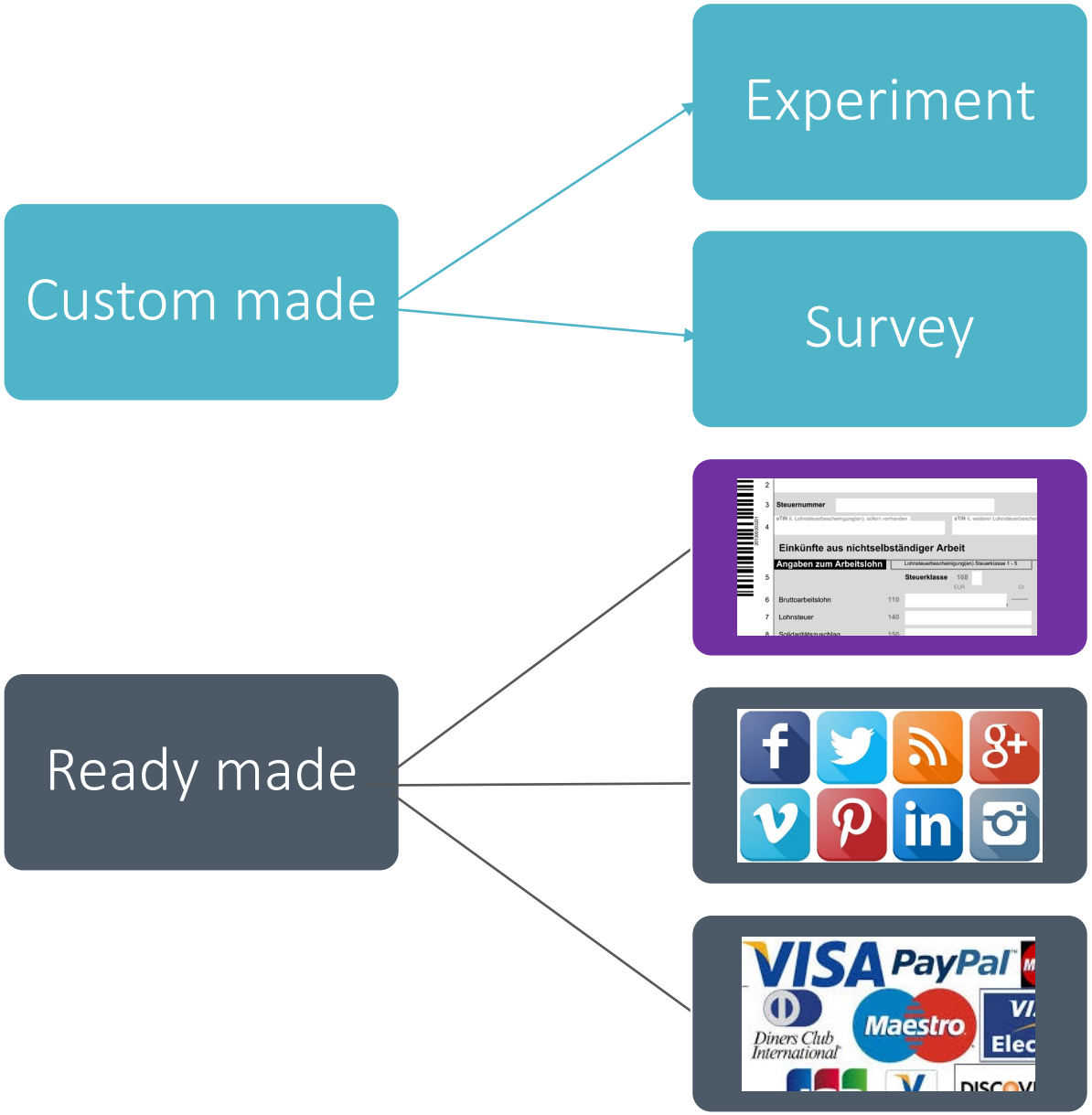
CITY OF BOULDER

The Hope


2
3 Steuernummer
eTIN lt. Lohnsteuerbescheinigung(en), sofern vorhanden
eTIN lt. weiterer Lohnsteuerbeschei
4
Einkünfte aus nichtselbständiger Arbeit
Angaben zum Arbeitslohn
Lohnsteuerbescheinigung(en) Steuerklasse 1 - 5
5 Steuerklasse 168
EUR Ct
6 Bruttoarbeitslohn 110
7 Lohnsteuer 140
8 Solidaritätszuschlag 150



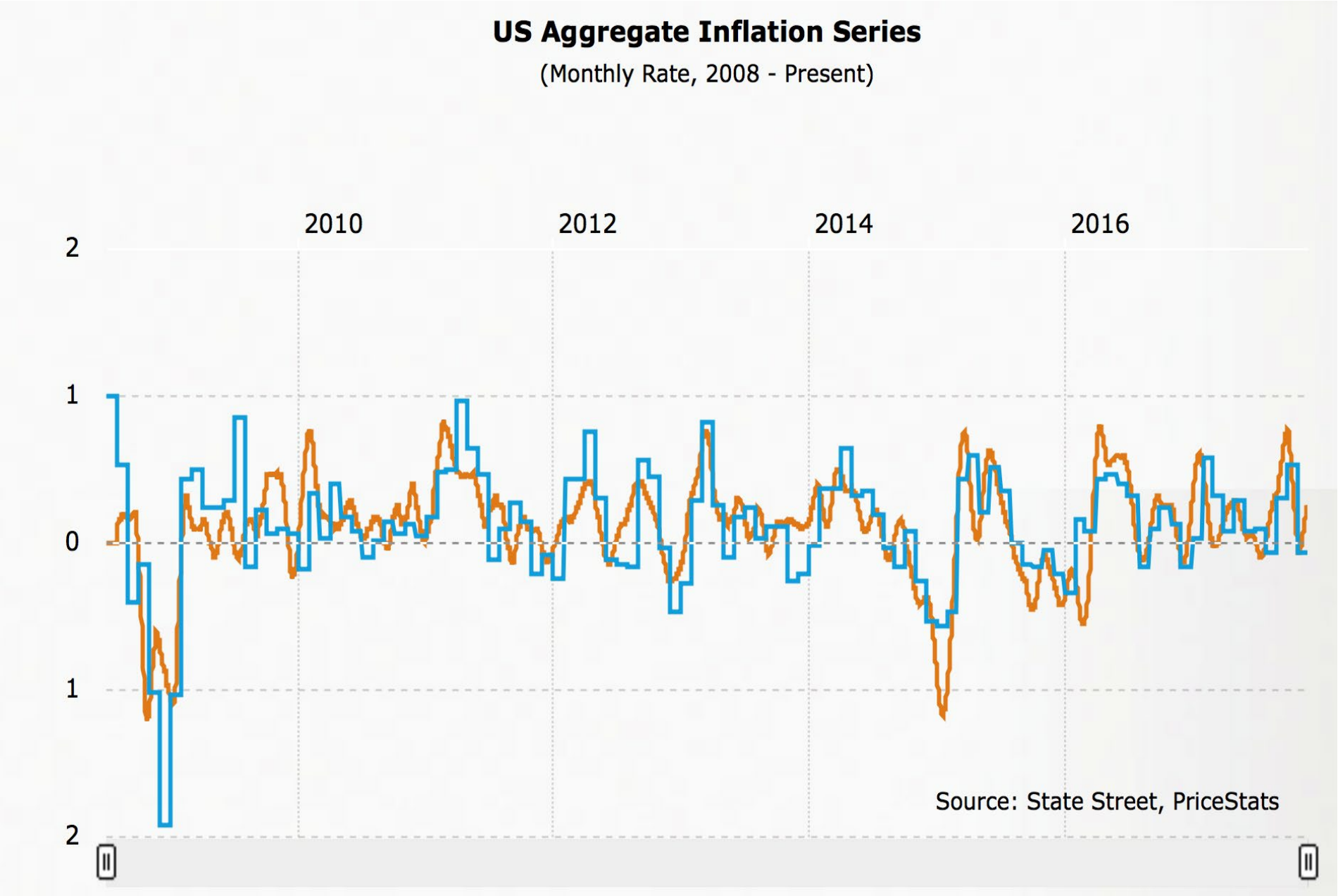
Groves 2011 -- https://www.census.gov/newsroom/blogs/director/2011/05/designed-data-and-organic-data.html



Salganik M. J. (2017): Bit by Bit. Social Research in the Digital Age. Princeton University Press

Economic Indicators

Examples for online data collection (and analysis)



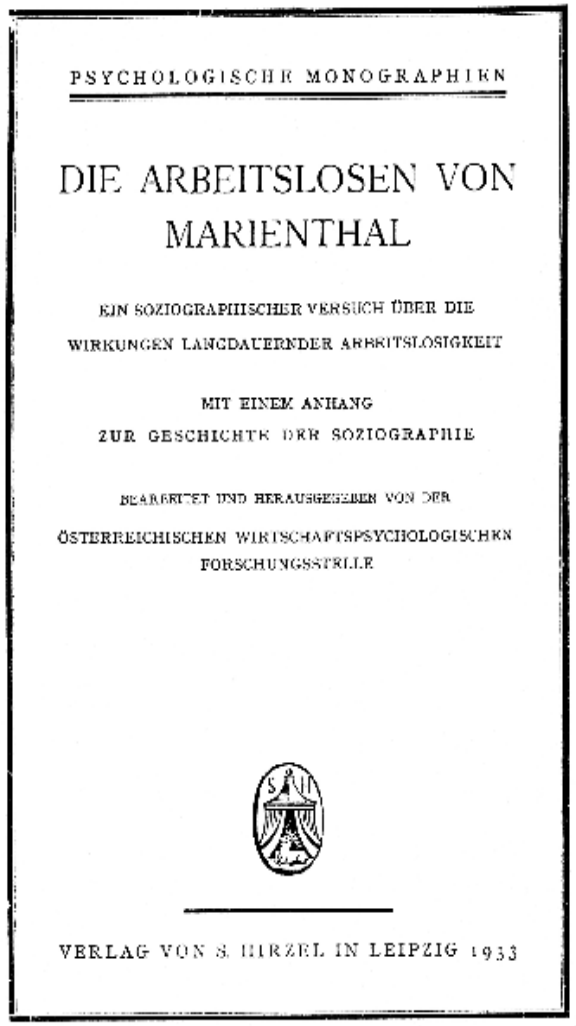
Official CPI

PriceStats Index

US Aggregated Inflation Series, Monthly Rate, PriceStats Index vs. Official CPI the PriceStats website 1/28/2018

Observations

GitHub


PSYCHOLOGISCHE MONOGRAPHIEN

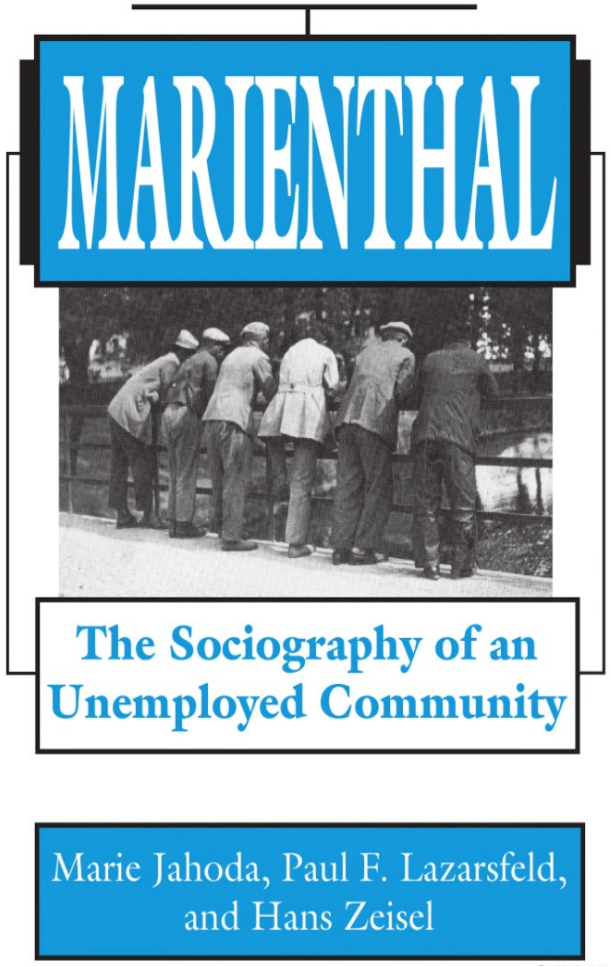
DIE ARBEITSLOSEN VON MARIENTHAL

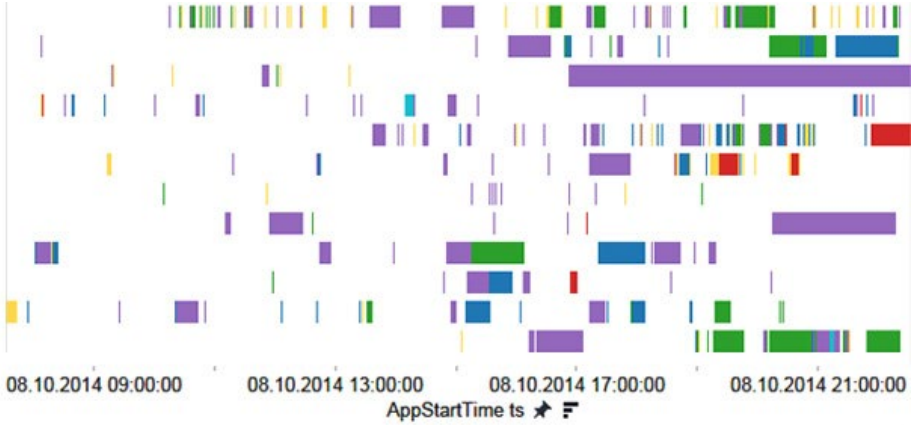
EIN SOZIOGRAPHISCHER VERSUCH ÜBER DIE WIRKUNGEN LANGDAUERNDER ARBEITSLOSIGKEIT

MIT EINEM ANHANG ZUR GESCHICHTE DER SOZIOGRAPHIE

BEARBEITET UND HERAUSGEGEBEN VON DER ÖSTERREICHISCHEN WIRTSCHAFTSPSYCHOLOGISCHEN FORSCHUNGSSTELLE

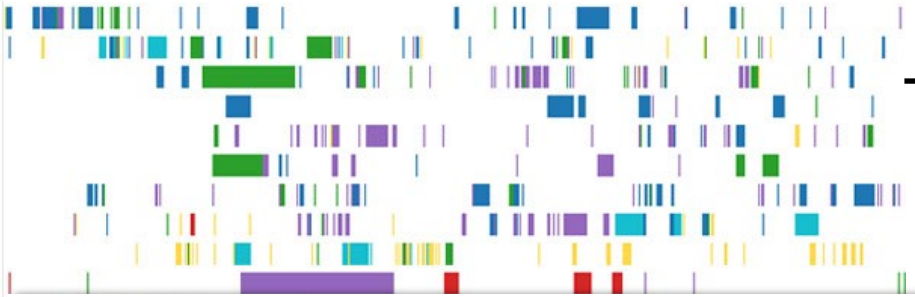
VERLAG VON S. HIRZEL IN LEIPZIG 1933



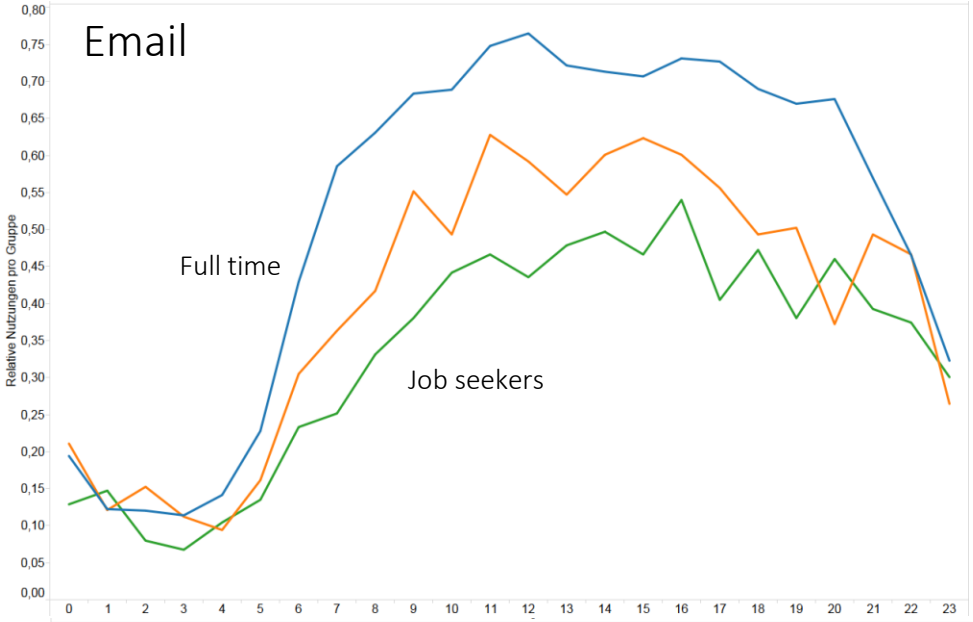


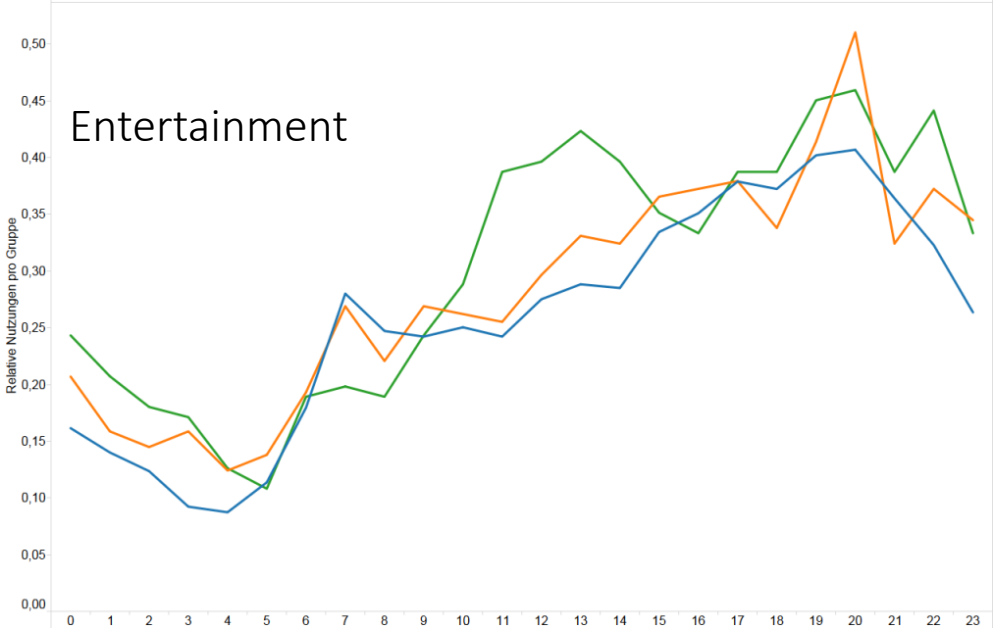
Full-time employed
→ App use past 5pm

Part-time employed
→ App use at noon



Job seekers
→ Continuous app use





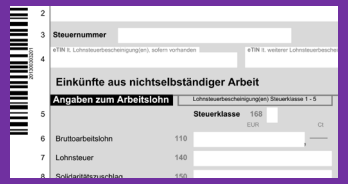
Schlickman et.al. (DataDiggers) App Nutzerverläufe statt Surveys? Eine Machbarkeitsstudie. DataFest Germany, Mannheim 2015, http://sswml.uni-mannheim.de/Teaching/DataFest%20Germany/DataFest%20Germany%202015/

However …

Work & $

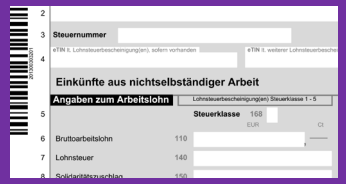
Experiment

Survey



Experiment

Survey



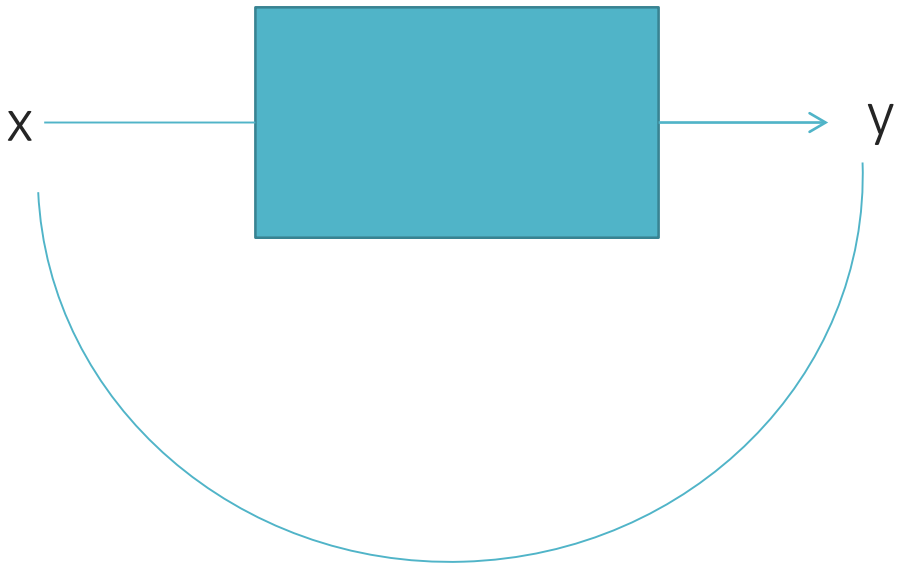
Work & $

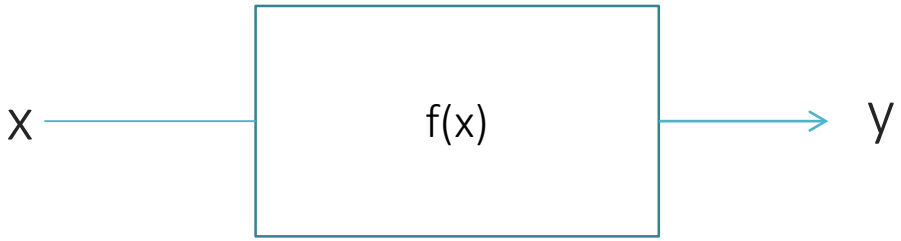
New Skills

Data Output/Access	Learn how to communicate results and distribute and store your data
Data Analysis	Learn a variety of analysis methods suited for different data types
Data Curation/Storage	Learn how to curate and manage data

Source: Usher in Japec et al 2015

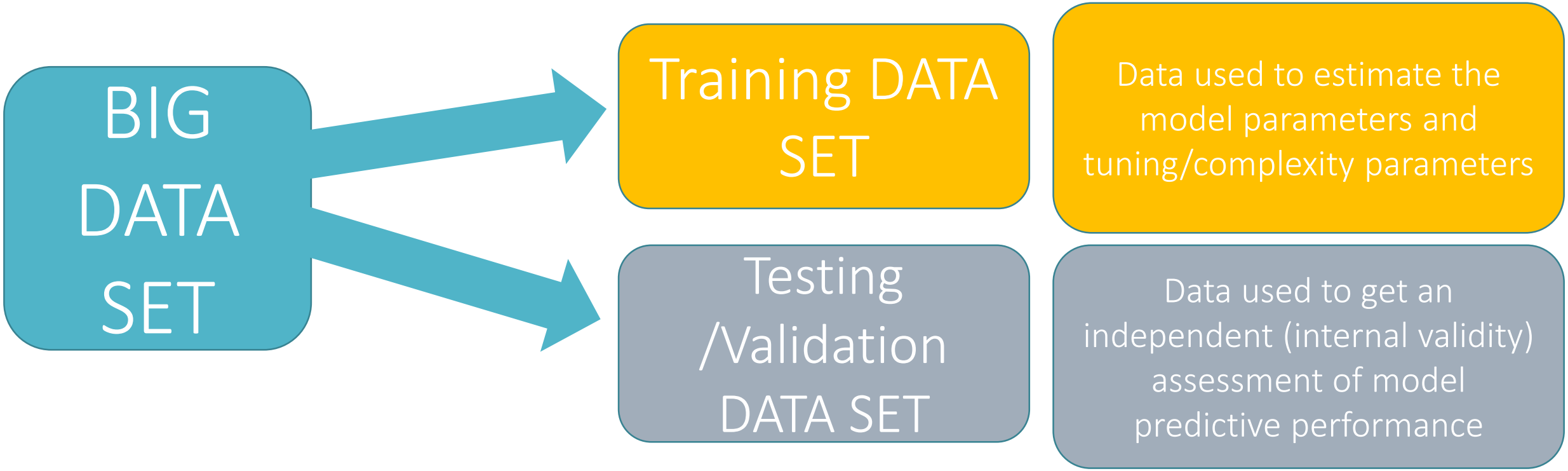
Machine Learning - AI





https://github.com/DataScienceSpecialization/courses from Roger Peng, Jeff Leek, Brian Caffo

Model Evaluation Strategy: Split Sample

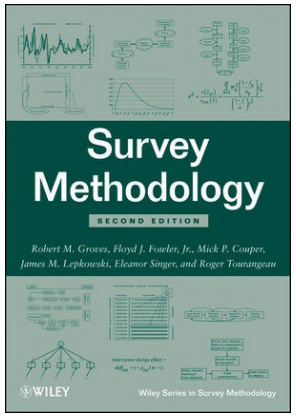


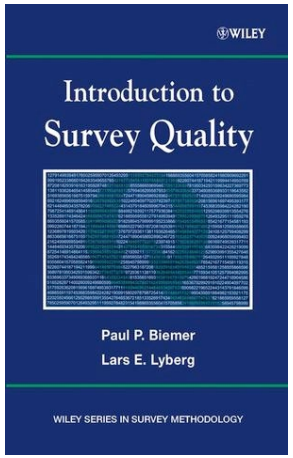
Risk in Learning from Data

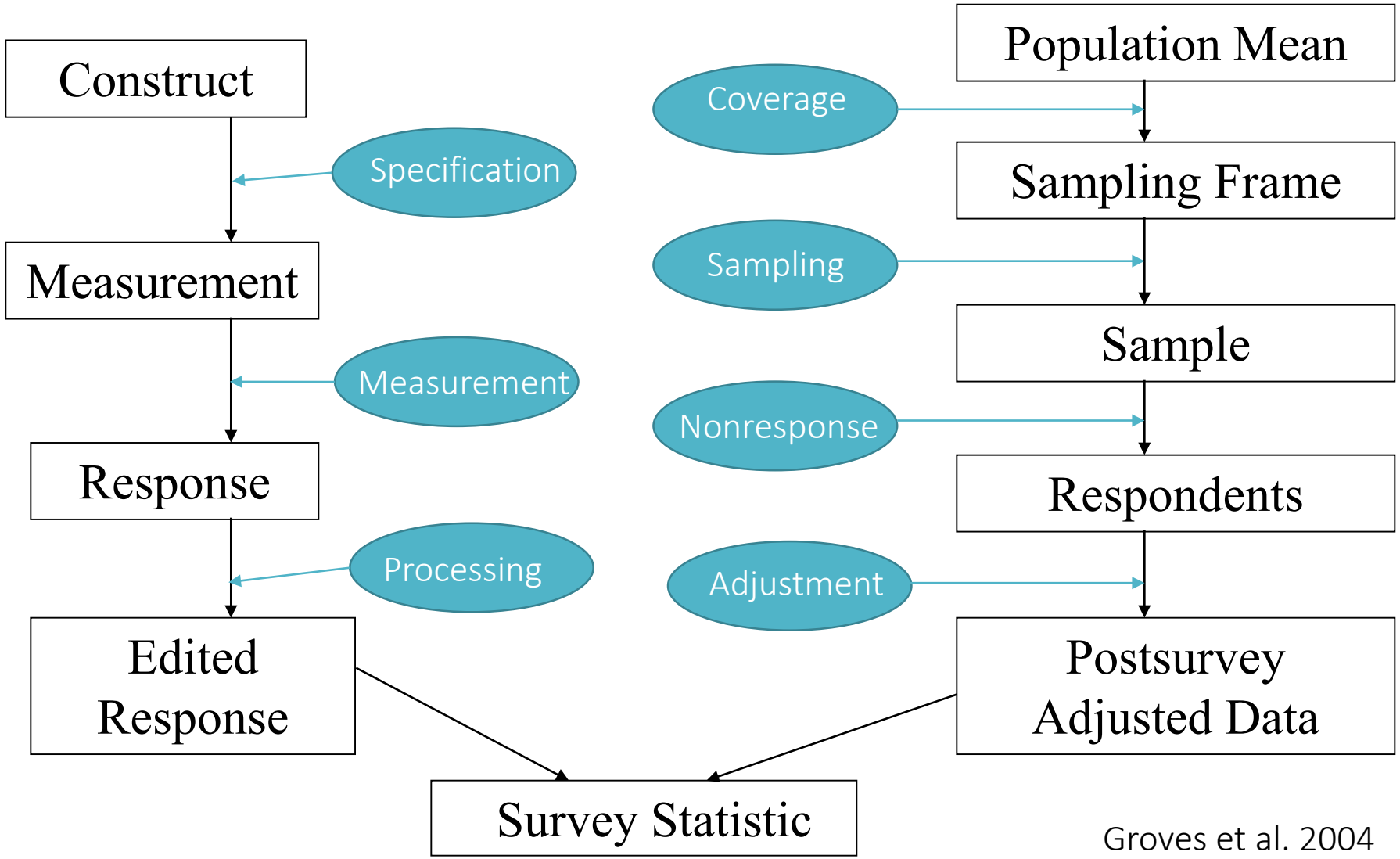
By far the worst approach is to wait for data quality problems to surface on their own.

T. Herzog, F. Scheuren, W. Winkler, 2007

Data Generating Process

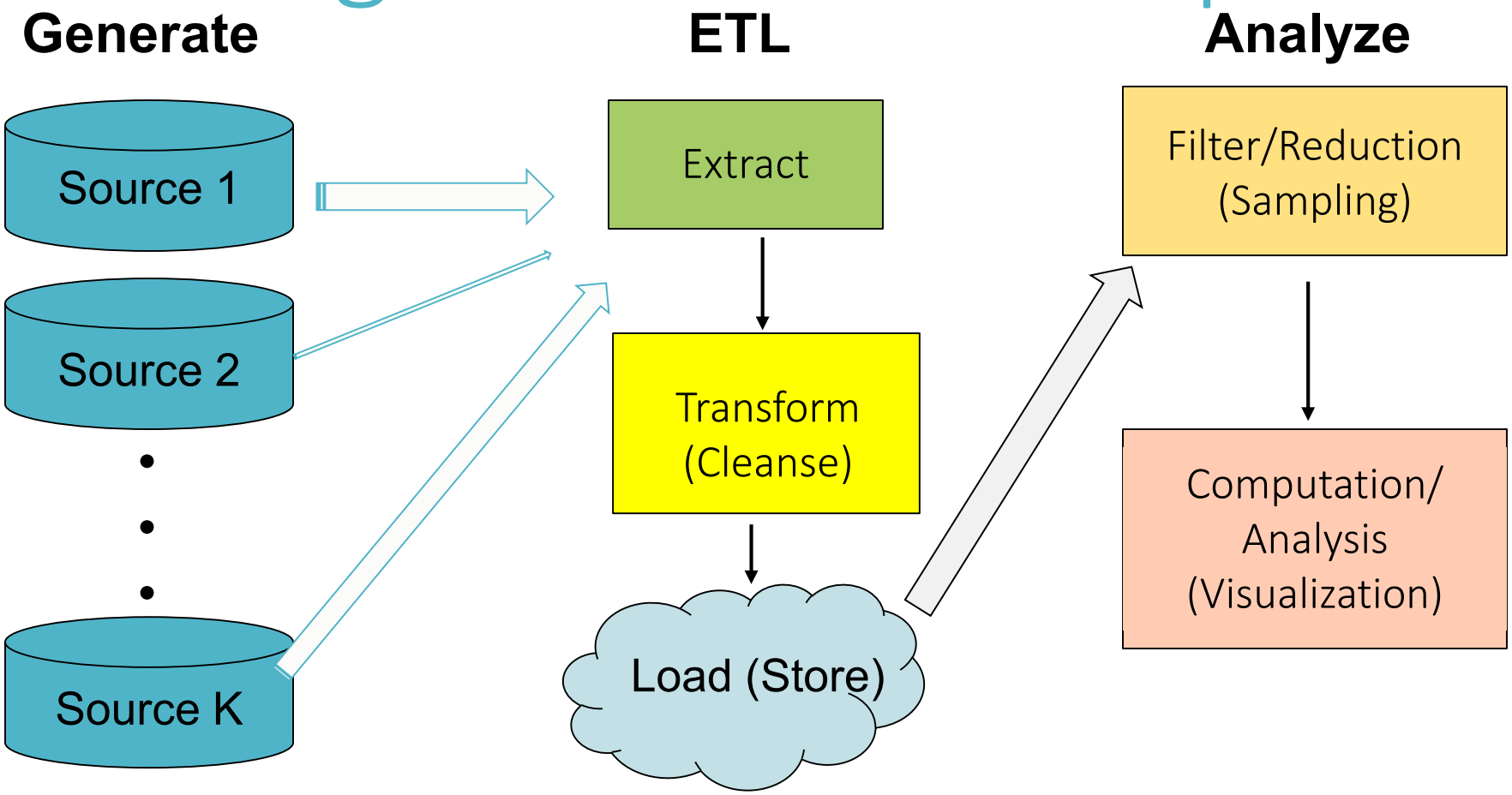






Groves et al. 2004

Big Data Process Map



Source: Paul Biemer in Japec, Kreuter et al. 2015 – AAPOR Task Force Report

Google Image Search: June 3rd 2018 Search Term: “University Professor”

DOMAIN EXPERT

User, analyst, or leaders with deep subject matter expertise related to the data, its appropriate use, and its limitations

SYS ADMIN

Team member responsible for defining and maintaining a computation infrastructure that enalbes large scale computation

METHODOLOGIST

Team member with experience applying formal research methods, including survey methodology and statistics

COMPUTER SCIENTIST

Technically skilled team member with education in computer programming and data processing technology

Big Data in survey research: AAPOR Task Force report. Japec, L.; Kreuter, F. ; Berg, M. ; Biemer, P. ; Decker, P.; Lampe, C.; Lane, J. ; O'Neil, C. ; Usher, A.. DOI: 10.1093/poq/nfv039. URL: http://poq.oxfordjournals.org/content/79/4/839.

Data Output/Access	Learn how to communicate results and distribute and store your data
Data Analysis	Learn a variety of analysis methods suited for different data types
Data Curation/Storage	Learn how to curate and manage data
Data Generating Process	Understand how to collect data yourself, and how data are generated through administrative and other processes.

Source: Usher in Japec et al 2015

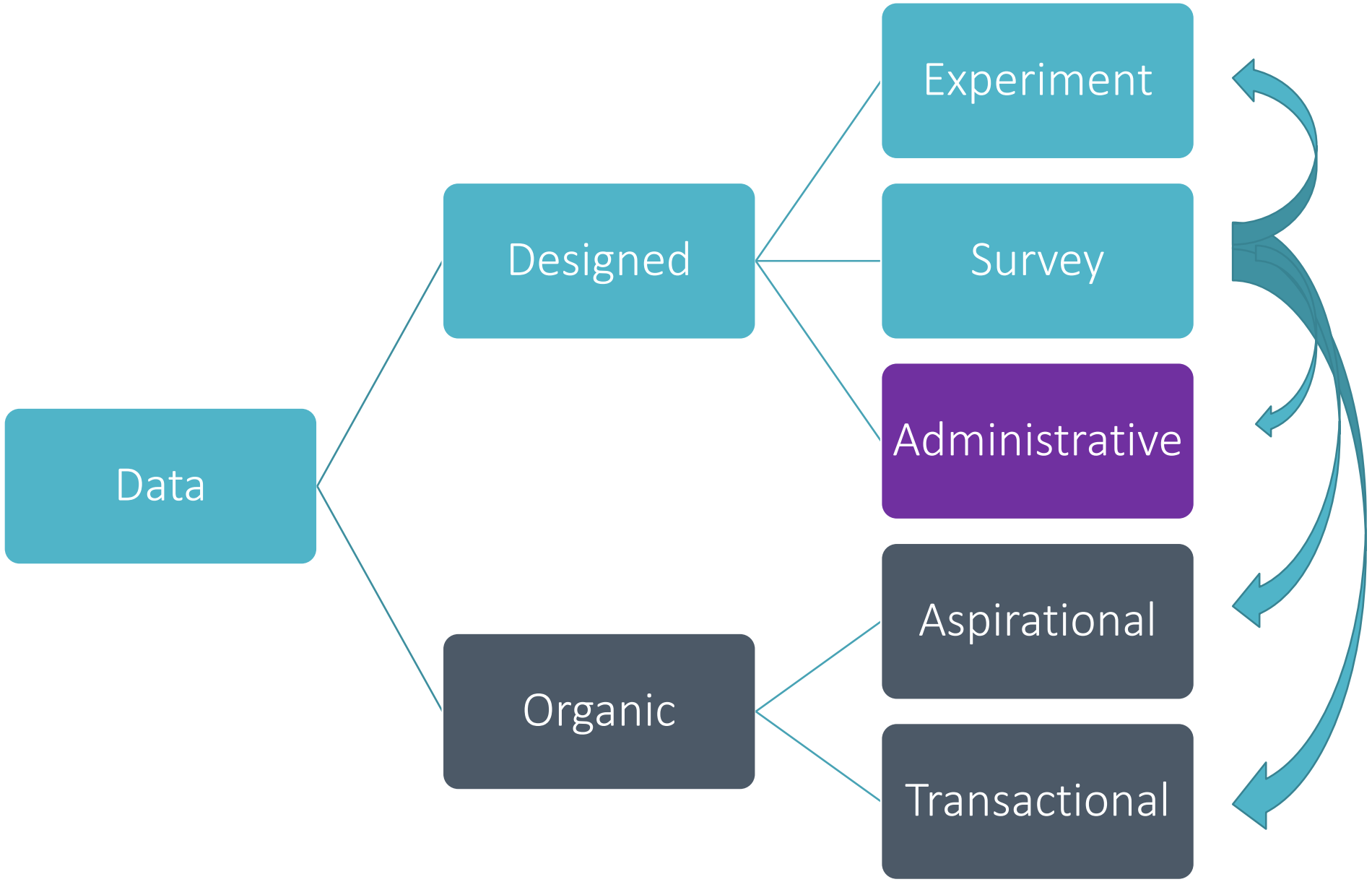
Data Output/Access	Learn how to communicate results and distribute and store your data
Data Analysis	Learn a variety of analysis methods suited for different data types
Data Curation/Storage	Learn how to curate and manage data
Data Generating Process	Understand how to collect data yourself, and how data are generated through administrative and other processes.
Research Question	Learn how to formulate your research goal and which data are best suited to achieve this goal.

Source: Usher in Japec et al 2015

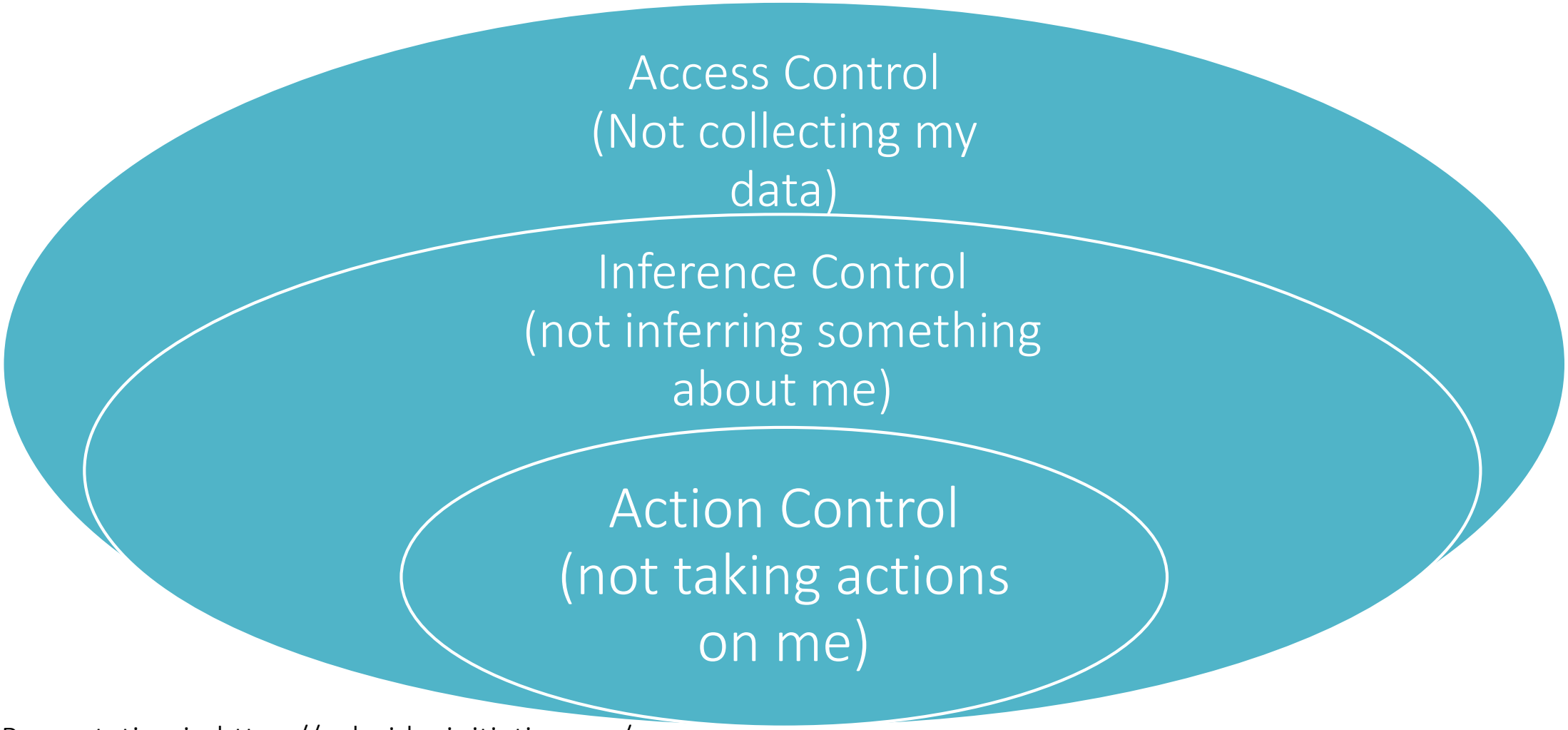
One way to think about a data analysis is to think of it as a product to be designed. […] Producing a useful product requires careful consideration of who will be using it.

Roger Peng, 2018

Risk in Combining Data


Data
Designed
Organic
Experiment
Survey
Administrative
Aspirational
Transactional

Consent to give up control



Ghani 2018: Presentation in https://coleridgeinitiative.org/

Front

The data you are about to provide to us would be much more valuable if you would allow us to link them with .... Do you agree to the linkage?

Back

The data you provided to us would be much more valuable if you would allow us to link them with .... Do you agree to the linkage?

Phone	Front	Back	Total n
% agree	90.8	78.7	598

Web	Front	Back	Total
% agree	82.6	62.4	520

Sakshaug et al. 2018

One Option

Introduction to International Program in Survey and Data Science

UNIVERSITY OF MANNHEIM
School of Social Sciences

JOINT PROGRAM IN SURVEY METHODOLOGY

SPONSORED BY THE

Federal Ministry of Education and Research

AUFSTIEG DURCH BILDUNG >>
OFFENE HOCHSCHULEN

International Faculty from Partner Universities

University of Mannheim

University of Maryland

University of Michigan

Manchester 1824
The University of Manchester

Utrecht University

LMU Ludwig-Maximilians-Universität München

Australian National University

UMass Boston

Pontificia Universidad Católica de Chile

International Faculty from the Industry

Google

United States Census Bureau

DEUTSCHE BUNDESBANK
EUROSYSTEM

facebook

BLS
BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR

Institute for Employment Research
The Research Institute of the Federal Employment Agency
IAB

RTI INTERNATIONAL

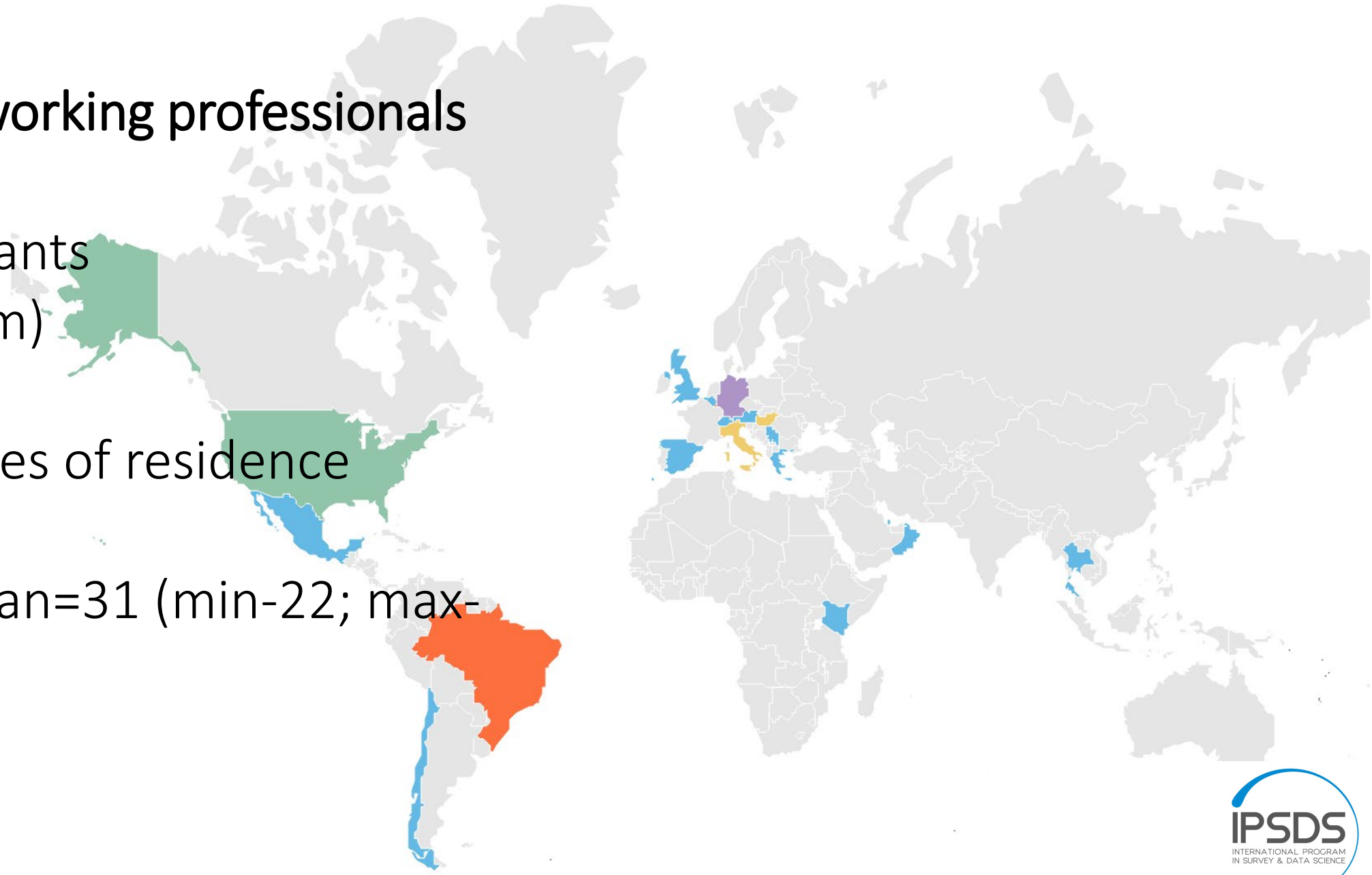
Westat

Abt SRBI

IPSDS
INTERNATIONAL PROGRAM IN SURVEY & DATA SCIENCE

4 IPSDS (Test) Cohorts

- **100% are working professionals**
- 47 Participants (27 f + 20 m)
- 19 countries of residence
- Age: median=31 (min-22; max-61)



IPSDS
INTERNATIONAL PROGRAM IN SURVEY & DATA SCIENCE

Selection of Courses

- Analysis of Complex Survey Data, 2 cr./4 ECTS
- Big Data and Machine Learning, 1 cr./2 ECTS
- Computer-Based Content Analysis I, 1 cr./2 ECTS
- Computer-Based Content Analysis II, 1 cr./2 ECTS
- Data Collection Methods, 3 cr./6 ECTS
- Experimental Design for Surveys, 2 cr./4 ECTS
- Fundamentals of Survey and Data Science, 3 credits/6 ECTS
- Generalized Linear Models, 2 cr./4 ECTS
- Inference from Complex Surveys, 2 cr./4 ECTS
- Introduction to Data Visualization, 1 cr./2 ECTS
- Introduction to Python and SQL, 1 cr./2 ECTS
- Introduction to Real World Data Management, 2 cr./4 ECTS
- Introduction to Small Area Estimation, 2 cr./4 ECTS
- Practical Tools for Sampling & Weighting , 2 cr./4 ECTS
- Privacy Law, 1 cr./2 ECTS
- Project Consulting, 6 cr./12 ECTS
- Questionnaire Design, 2 cr./4 ECTS
- Review of Statistical Concepts (bridge course)
- Web Survey Methodology, 2 cr./4 ECTS

IPSDS
INTERNATIONAL PROGRAM IN SURVEY & DATA SCIENCE

Open Enrollment Couress - 2019/2020

- Applied Sampling (Sampling I), 2 credits/4 ECTS
- Data Confidentiality and Statistical Disclosure Control, 2 credits/4 ECTS
- Introduction to Record Linkage with Big Data Application, 1 credit/2 ECTS
- Item Nonresponse and Imputation, 1 credit/2 ECTS
- Measurement Error Models, 1 credit/2 ECTS
- Usability Testing for Survey Research, 1 credit/2 ECTS
- Web Scraping and API, 1 credit/2 ECTS

IPSDS
INTERNATIONAL PROGRAM IN SURVEY & DATA SCIENCE

Flexible & engaging online learning environment

- Online learning environment accessible from anywhere in the world (taught in English)

- Small virtual classroom with a mix of synchronous & asynchronous learning
  - *Pre-recorded lectures split into small video units*
  - *Required readings and (bi)weekly assignments*
  - *Discussion forums*
  - *Weekly online meetings*

  } 8-10h per week

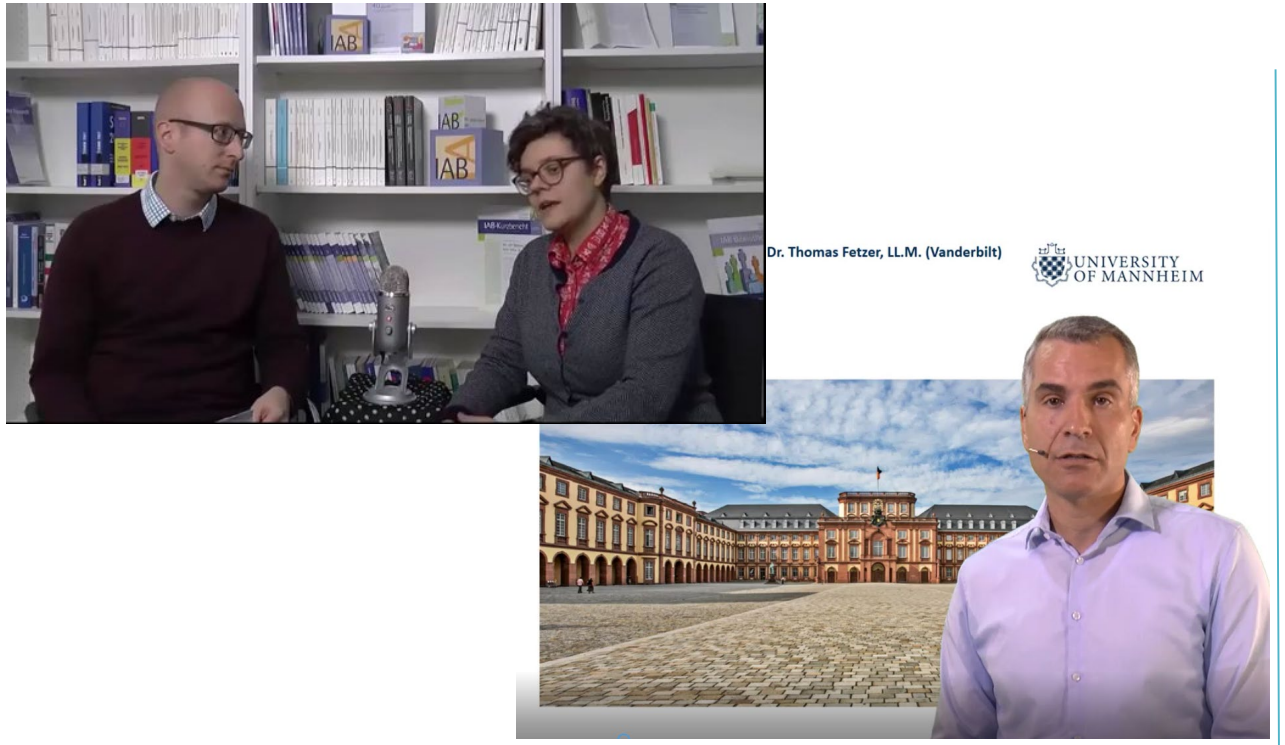
- Annual on-site networking activity with fellow students from five continents

- Wide variety of options: from individual courses or course sequences to a modular program

IPSDS
INTERNATIONAL PROGRAM IN SURVEY & DATA SCIENCE

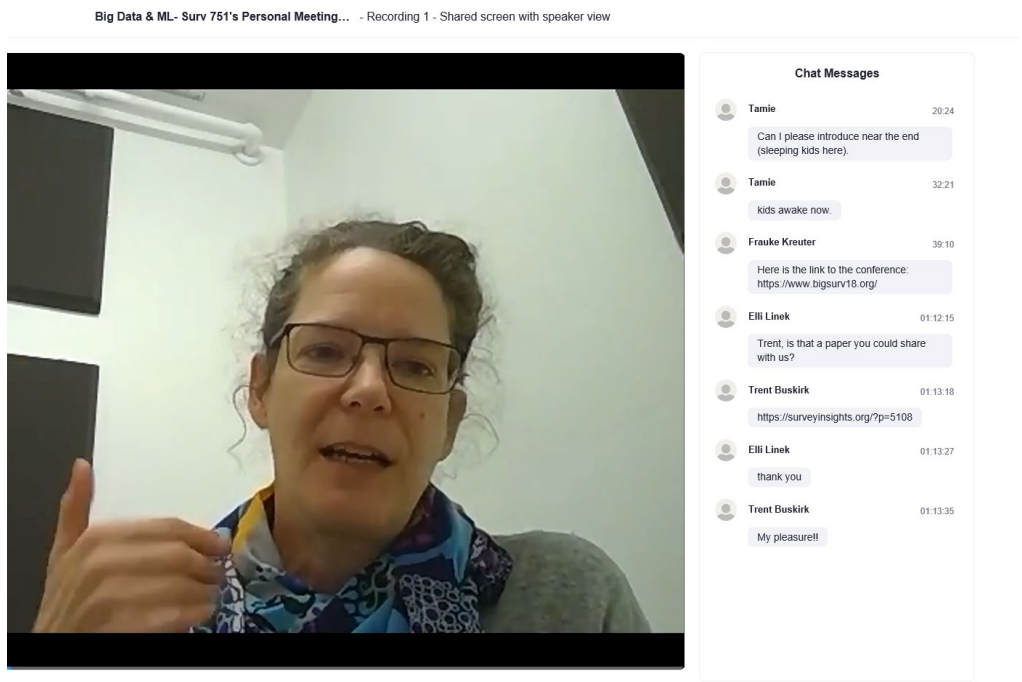
Format

Asynchronous



- Pre-recorded lectures (split into small video units)
- Required readings and (bi)weekly assignments
- Discussion forums

Synchronous



- Small virtual classrooms
- Weekly 50-minute discussions led by the instructor

Summary

1. Data are plentiful and increasingly accessible
2. Costs of statistic production often not reduced but shifted from collection to processing
3. New data need new skills. Those can be acquired by existing staff, manageable by social scientists and content experts. Best to work in teams
4. Mindset shift to focus on quality (as it is strength of UN) is even more important
5. Ethical and cognitive psychology consideration important when (asking for) combining data, to stay within a contextual integrity of data use

THANK YOU!

fkreuter@umd.edu